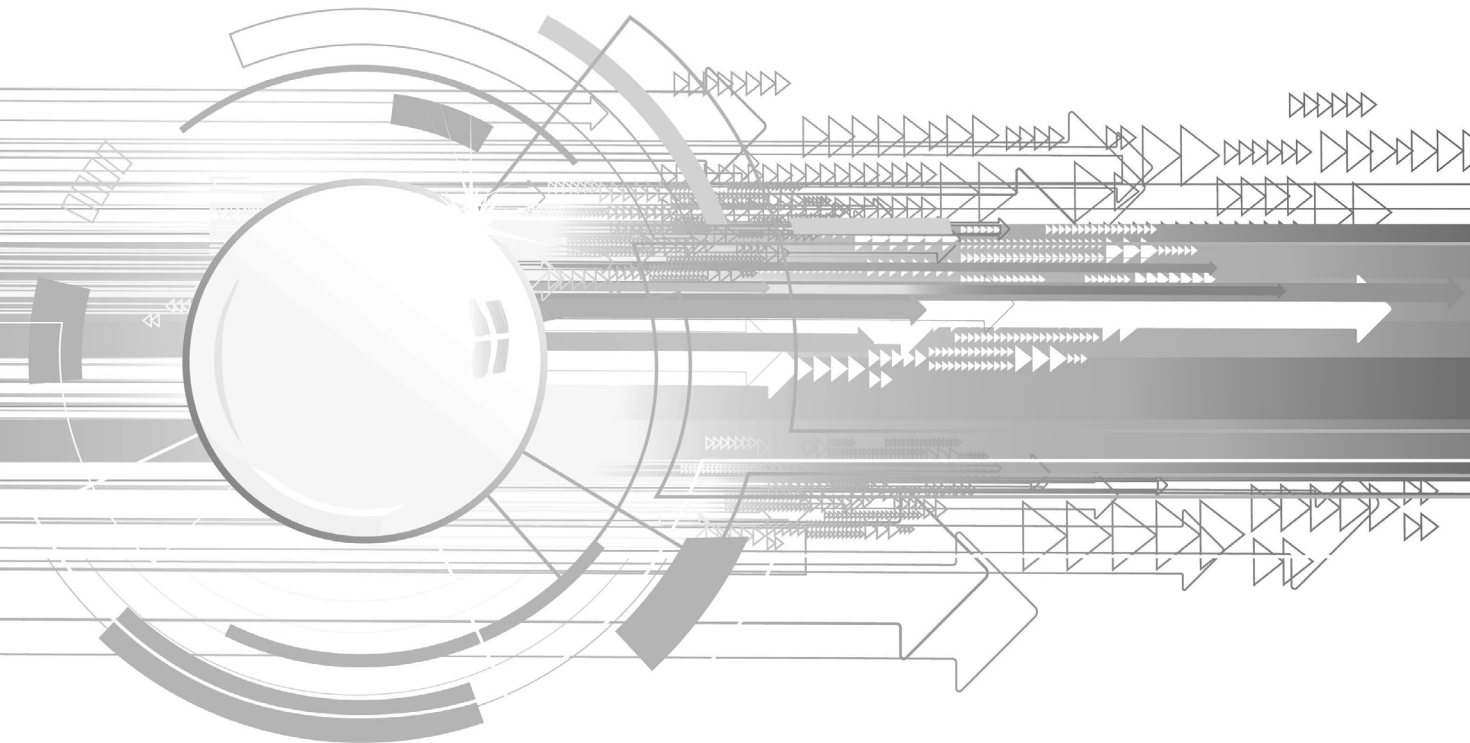


PART



Preparation

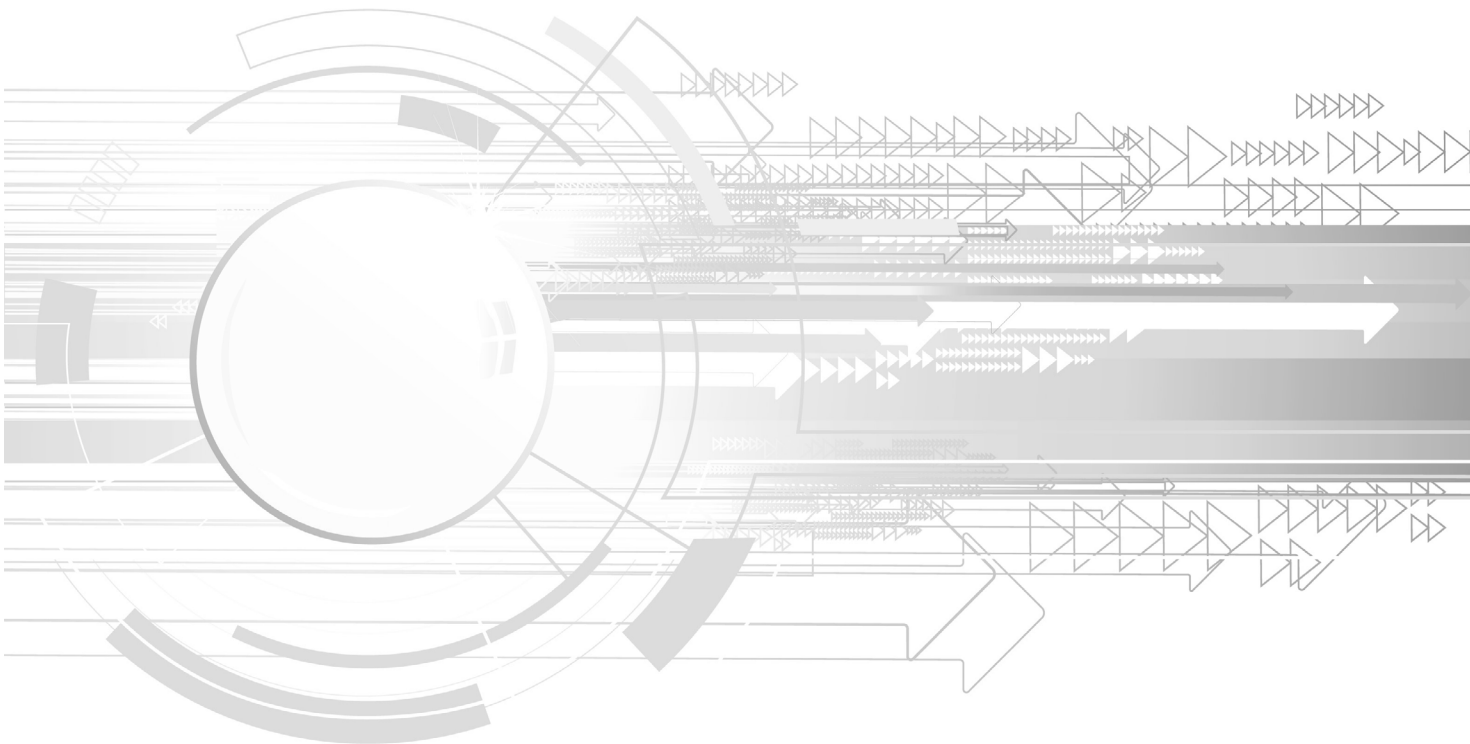
A management overview discussion on data warehouse and business intelligence systems.



CHAPTER

1

Data Warehouse and Business Intelligence Overview



Many organizations are contemplating, developing, planning, or currently using a data warehouse system. This type of system is becoming more popular as organizations are strategically deciding to treat their corporate data as an asset. Companies are funding projects to structure, organize, cleanse, document, and centralize their data. The depth of understanding derived from these complex systems provides companies with a discernible competitive advantage in the marketplace.

Business and information technology are interwoven. In today's organizations, one cannot function without the other; therefore, they use best practices to guide development efforts for all projects. For a moment, consider the complexities facing IT and the direction and requirement of the business. For most businesses, IT can magnify the business efficiencies if done right, or it can devastate the business resources if done wrong.

This book will define and explain the many facets of a data warehouse system from practical business and technical perspectives. In short, this book gives insights into data warehouse and business intelligence systems with a focus on best practices and a goal of success for the business and IT alike.

Business Intelligence Overview

Empowering business decision makers with trusted data in a value-driven context that is usable and delivered in a timely manner is business intelligence (BI) in a nutshell.

Business intelligence is a wide topic. Many books have been written on business intelligence detailing purpose, value, and very specific usage scenarios. This topic is fundamental to a data warehouse from a usage point of view. BI is the part where the business uses the underlying data to support informed business decision making and processes.

Let us start with BI first to get an appreciation for the basics, and then we can move into a larger data warehouse perspective.

Definition

Business intelligence is an umbrella term referring to skills, processes, technologies, applications, and practices used to support business decision making. Business intelligence deals with used data, or past data, in a desired context to help make business decisions for tomorrow.

Business intelligence is mostly focused on internal information about operational issues as they pertain to tactical and strategic planning. Information is typically structured in one fashion or another and is focused and/or gathered from current

business processes. The output of the underlying data is mostly used for internal analysis but can also be used in conjunction with external analysis such as with SWOT and PEST competitive analysis efforts.

SWOT analysis provides information helpful in matching a firm's resources and capabilities to the competitive environment in which it operates. A scan of the internal and external environment is an important part of the strategic planning process. Environmental factors internal to the firm usually can be classified as strengths (S) or weaknesses (W), and those external to the firm can be classified as opportunities (O) or threats (T). Such an analysis of the strategic environment is referred to as a SWOT analysis; see QuickMBA: <http://www.quickmba.com/strategy/swot/>. PEST (Political, Economic, Social, and Technological), is a scan of the external macro-environment in which the firm operates; see QuickMBA: <http://www.quickmba.com/strategy/pest/>.

Business intelligence is based on initial key performance indicators (KPIs) in the quest for determining business goals. These KPIs can be broken down into measures, aka facts, fundamentally sourced from within the organizations' operational systems or aggregated from these fundamental operational measures. Facts are normally numeric and quantifiable, thus being additive. Typically, analysis in a BI environment is performed at an aggregated level rather than at an instance level; for example, an organization wants to know how many customers exist, live, and/or purchase in a specific region as opposed to where an individual person lives and how much he or she purchased. The later scenario can be taken or derived from an existing operational system, but, that said, there is no reason why this information cannot be extracted from a business intelligence system as long as the system holds data at the appropriate granularity. I have witnessed situations in several organizations where the fundamental data of how much an individual customer has purchased is not always available in the operational systems, and therefore the BI environment is used as a central merging area of events and fundamental data. Some may create an operational data store (ODS) for current granular information, but an ODS tends to have limited history, which is why a BI environment is most often used.

A good business intelligence system has been described as being accurate, timely, of high value, and actionable: *accurate* in the sense that data is trusted; *timely* meaning that data is available on a regular schedule; *of high value* meaning useful to the business user; and *actionable* meaning that the information can be used in the business decision process. Concluding that the organization has 12 power-lift trucks is nice, but is it useful to the business decision process? If it is, then there is a purpose to having the underlying data; if not, what is the point of spending the effort and budget on baking the data when it is not useful to the business or actionable by the business?

Value of Business Intelligence

Business intelligence gives the business decision makers the ability to query the data themselves.

Back in the late 1970s when I started in the computer industry, when a business user had a question requiring access to the data, they would send a memo via interoffice mail requesting a report to the IT department. The memo would explain the required specifics and the IT manager would schedule someone to work on the effort. The mini-project involved lots of telephone calls back and forth with the business as both tried to figure out exactly what was required along with the underlying data components. Once the purpose and requirements were more or less understood, IT would write a program to create the necessary report. The process took days and sometimes weeks depending on the complexity of the request. The effort was always hampered due to communications. IT was trying to figure out what the user wanted in data terms records and fields (no tables and column terminology back then), and the business user was trying to figure out how to explain the request using as simple and understandable business terminology as possible. IT typically had no clue as to the data context or what decisions the user was trying to formulate by using the report. Conversely, the business user had no clue as to the records and fields holding the data, which resulted in neither IT nor the business user knowing exactly nor effectively how to guide the other. Eventually, as IT learned more of the business and the user more of the data, together business and IT were able to produce reports in a timely manner of a day or two.

Today business intelligence empowers the business user to query the data directly. This seemingly small advancement has helped reduce the decision-making process from days and possibly weeks down to minutes. BI is to the business decision process as a television remote control is to a couch potato—empowering. Imagine how long it would take (or how long it used to take) to get up and walk to the television set each time you wanted to change channels! Having a remote control massively reduces this time and effort while empowering the viewer to channel-surf at will. The same empowerment concept when applied to the computing industry is called “business intelligence.”

Business intelligence massively increases the business users’ ability to process information. Just as channel surfing allows the viewer to get an idea as to the actual content of what is available on the many television stations fairly quickly, business intelligence gives the user the ability to query the information directly and as often as desired without the need to involve IT in every step. Of course, having a requirement focus in advance allows IT to fine-tune efforts on designing and building an environment to support this efficiency with only the required data for the business needs at hand, just as a television guide allows the viewer to quickly search

for programs on specific days and at specific hours, and using specific themes such as comedy, drama, movies, and so forth.

Due to the quick turnaround between data availability and business usage of the data, feedback of data quality issues has led the proverbial IT construction and repair crew to react much more rapidly to evolving business needs. This has led to data being available in a structured and timely manner with much focus on accuracy and quality, allowing the business to gain more insight and react more quickly to the business environment.

In the ideal, business intelligence is

- ▶ Empowerment—directly usable
- ▶ Fast—responsive
- ▶ Timely—available
- ▶ Accurate—trusted with quality
- ▶ Usable—has value

Breakdown of Business and Intelligence

The term business intelligence implies that the business has a person or group of people capable of making business decisions, and that the underlying information upon which the decision is based is trusted.

The people involved in decision making hopefully are in a position whereby they can be expected to have the background, education, and experience to make the required decisions to move the business forward, whether in a functional, tactical, or strategic manner. These people need a stone to stand on when deciding which direction the organization, line of business, department, or resulting action should take. This proverbial stone is information. Without proper or trusted information, the decision process can easily be contaminated, thus rendering the final outcome completely inappropriate.

Information is data in context. For information to be correct, a better word would be *dependable*; the underlying data must be trusted. Therefore, great importance is placed on the quality of the data. If management receives multiple conflicting informational details such as counts of products sold or cost of operating production-line machines, how can a person derive appropriate and proper decisions and therefore correctly know which business direction to take? If you drive to a fork in the road with no signs, how can you tell which route to follow? A sign must be posted pointing the way; this is fundamentally trusted information in a timely manner.

Data is fundamental to information, and this data must be trusted. To be trusted, the data must have a high level of reliability, aka quality, otherwise known as integrity.

Several years back, a project undertaking involved an automobile insurance organization that was trying to formulate a marketing strategy based on the types of automobiles they insured. Unfortunately their data-capturing system for automobile types, makes, and models was completely manual and did not deliver appropriate data integrity. Company clerks, who manually typed the individual insurance policy details into the computer systems, would simply type in whatever the policy stated, with no quality control. Unfortunately, to make matters worse, at times there were many typing errors that went uncorrected. For instance, a FORD automobile was entered as FORD, FRD, ORD, RFOD, Mustang Ford, 98 Food and so forth. Without cleansing this data at the source to ensure that the policy was indeed describing a FORD automobile, every use of this data item down the line within this organization was at risk. Absolutely no decisions based on automobile types, makes, and models would be accurate, and therefore this information was completely lost to the organization. Imagine an automobile insurance company not being able to accurately describe automobile manufacturers for their policies—seems silly, doesn't it? Well, this is a true case that went on for years. I am happy to report that since the organization initiated a business intelligence effort along with a data quality effort, the data is now cleansed, trusted, and used regularly within the business. Now the brokers, underwriters, and adjudicators can all trust the data and make informed decisions upon this aspect of the business data.

Business Intelligence Success Factors

Many of the factors that affect a successful BI implementation are not necessarily the same as those for a data warehouse. A data warehouse system typically expresses itself as BI to the end user, but a data warehouse system can or may focus only on the data aspect without specific regard for business usage; in such a situation, business intelligence becomes a subcomponent of the larger data warehouse system. A data warehouse that is not imminently and directly usable to the business may be part of a master data management effort to centralize on a specific vocabulary or to coordinate a centralized data effort. This effort would clearly be a foundation phase preceding business usage—getting all the ducks in a row, so to speak!

As noted in Wikipedia (“Business Intelligence,” 2010), Naveen K. Vodapalli lists the following as the critical success factors for a business intelligence implementation:

- ▶ Business-driven methodology and project management
- ▶ Clear vision and planning
- ▶ Committed management support and sponsorship
- ▶ Data management and quality issues

- ▶ Mapping the solutions to the user requirements
- ▶ Performance considerations of the BI system
- ▶ Robust and extensible framework

It suffices to say that a BI effort must be business-driven to be successful. The nature of BI is to add value to the business; therefore, there must be a clearly focused business purpose driven by the business itself and championed by executive management so that it does not turn out to be some manager's pet project or hampered by internal politics.

Since the underlying data is fundamental to the success of the effort, there must be a tight collaborative partnership among IT, the developers, and the business unit—the customer. IT must put heavy thought into data management and requires executive management to enforce any source-system data quality corrections if need be (and there is always a need!!). Additionally, IT must perform due diligence in ensuring that whatever is built is flexible enough to be extended with future phases and efforts.

In the end, the business will expect a realistic response time from the BI solution, whether this expectation is verbalized or not. If the proverbial button is pressed and the system takes 30 minutes to respond with a resulting report, there will be issues, the worst being non-use of the system. If the business does not use the newly built solution even though it cost thousands or millions to create, the overall effort and investment are lost. Therefore, ensuring a usable environment is a priority, and such usage is a sign of success for both the business and IT.

Purpose of BI

There are many purposes for and methods of performing business intelligence. In other words, there are many types of business intelligence and business analysis. Understanding each type and therefore being able to plan for the right environment can add much value to the business.

Every organization produces some sort of reporting regarding the ongoing aspects of their business. Management spends much time reading, interpreting, and basing decisions upon these reports. Otherwise, how would individuals know the current status of events within the business? At a rudimentary level, enterprises capture inventory, production progress, sales, and so forth. To track events in such areas requires a report of one sort or another. As the organization grows and events unfold, comparisons to previous metrics and efforts are analyzed to know if progress is above, below, or the same as previous time periods. Viewing the data from different perspectives such as by product types, by geographic regions, or whatever adds more clarity to events. More information gives greater insights, which allows for a clearer picture of the current environment with expectations of what to do next. Business

intelligence depends on business data being available and usable by the management to gain business insights to support business decision making.

Years ago I was analyzing the business intelligence requirements of a large European bank. They were all excited about creating a data warehouse system and producing great business intelligence to guide their efforts forward. The number one point on their list was to ensure a business-as-usual environment. This meant that their first priority was to replicate reports currently being used by the business. The idea was to add value to the reporting environment by setting a new foundation platform whereby there would be a one-stop-shopping scenario. All data would come from this one central and trusted environment. This meant that all reports would be based on the same data—no more disparate systems, just one central warehouse holding cleansed data. From this foundation, advancements could be planned, designed, and constructed to move the organization forward. All seemed well thought-out. So we proceeded to get a view of their reporting needs. At that point someone literally wheeled in 120,000 business reports that had been produced from hundreds of source systems. Imagine running a business on that many reports produced from so many different systems. Imagine the confusion, the discrepancies, the redundancy, the data quality issues, and the quarterly reporting craze that must have been going on.

The point is that without analyzing the purpose and type of business intelligence, organizations are bound to create and re-create their reporting environment over and over again, basing their future on the exact same errors as in the past. Determining the type of business intelligence to be done and the output delivery method will greatly enhance the overall usage and effort.

For the bank scenario I described, there was obviously more work to be done in classifying the different lines of business; determining the business processes, areas of analysis, methods and priority of analysis; understanding who was to access what; whether drill-down and drill-up was required; how dynamic the reports were to be, and so forth. Much more analysis needed to be done before any planning or design efforts could begin.

Business intelligence has many purposes, including but not limited to:

- ▶ Benchmarking or baselining
- ▶ Trending or predictive analysis
- ▶ Affinity grouping, aka market basket analysis, or segmentation
- ▶ Performance management
- ▶ Associative analysis, aka data mining
- ▶ Subject area analysis

Each of these serve a specific purpose based on usage. Baselining refers to creating an environment whereby, for instance, monthly store sales can be compared

in terms of being more or less than the previous 18-month rolling average store sales, globally, regionally, or locally. Predictive efforts could involve the analysis of expected sales over the next three to five years given the previous several years of monthly sales while also being compared to its associated industry sector sales. Affinity grouping could help marketing in understanding the top 100 selling items as they relate to the top ten secondary selling items, perhaps helping product and store planogram efforts. A specific focus on customer relationship management may look at customer lifetime value or customer lifetime progression analysis, tagging products or services to customers as they age or enter lifecycle changes. Data mining could include the manual effort of discovering associations between data components in the quest to analyze market segments as they pertain to events. Subject area analysis may involve insights into product lines.

The point here is to understand the purpose of the business analysis and to create business intelligence based on this specific purpose. Keeping a distinct business requirement focus while managing the fundamental underlying data is the key to a successful and flexible business intelligence solution.

BI User Presentation

The business intelligence outcome or presentation of the outcome can be in several forms:

- ▶ Reports
- ▶ Queries
- ▶ OLAP
- ▶ Dashboards
- ▶ Scorecards

Everyone is familiar with reports. These are static, typically scheduled pre-run routines that produce specific layouts. You get what you see. Organizations have been using these since the beginning of time.

When a person desires to look into specific correlations or details, there is always the option to write structured queries either manually or with assistance. These are typically SQL-based but can be assisted with the aid of the drag-and-drop features of BI tools. An example of a structured query is: `Select Product.Name from Product where Product.Color = "blue"`.

Online analytical processing (OLAP) is another form of data inquiry mechanism. This method gives dynamic aspects to typically static reporting. As the term denotes, these reports are online as opposed to being created in a batch run or printed for manual delivery.

OLAP empowers the end user with the ability to actively drill down or drill up within a report. A starting point might be sales in a region such as Canada for a specific time period such as 2010. The user may then wish to look at (drill down to) sales per province and then down again to sales by major city. Essentially OLAP is the concentration of many reports all rolled into one.

In the old days, the following reports could have been produced independently:

- ▶ Sales in Canada for 2007: 1 report
- ▶ Sales in Canada per Q1, Q2, Q3, Q4 for 2007: 4 reports
- ▶ Sales in Canada per month for 2007: 12 reports
- ▶ Sales in ten Canadian provinces for 2007: 10 reports
- ▶ Sales in ten Canadian provinces per Q1, Q2, Q3, Q4 for 2007: 40 reports
- ▶ Sales in ten Canadian provinces per month for 2007: 120 reports
- ▶ Sales by ten Canadian provinces per month for 2007 by major city: $10 \times 12 \times 20 = 2400$ reports

Total number of reports calculated the old way would be: $1+4+12+10+40+120+2400 = 2587$ reports. Now add five years of history, and that's an easy 12,935 business reports. Now add 20 products, and that brings the total to 258,700 possible reports.

We could of course produce the final detailed report as seen later in this chapter in Figure 1-5, but in many instances this is the result of multiple request iterations over a number of months; “could one more item be added to the report please?” I remember spending lots of time in the early 1980s rewriting report programs, rescheduling nightly batch jobs, and rediscussing the produced report with “what-if” we added this or that type of scenarios. The more dimensions added, each with its own cardinalities, the more complicated the report became and the longer it took to generate.

The OLAP method requires an underlying structure typically designed in a cube fashion, which allows for the dynamic creation of reports based on the grain of the dimensions. *Granularity* refers to the level of detail; date granularity can be day, week, month, quarter, year, and so forth. These dimensions or varying parameters, for reporting scenarios just described, are

- ▶ Date hierarchy: time, time period, or date: calendar year, calendar quarter, calendar month
- ▶ Geographic hierarchy: geographic area: country, province, major city
- ▶ Products (which may have several hierarchies in its own right)

An example of a typical OLAP cube is Figure 1-1. The little squares are the representation of a measure (“sales” for our example) by specific product,

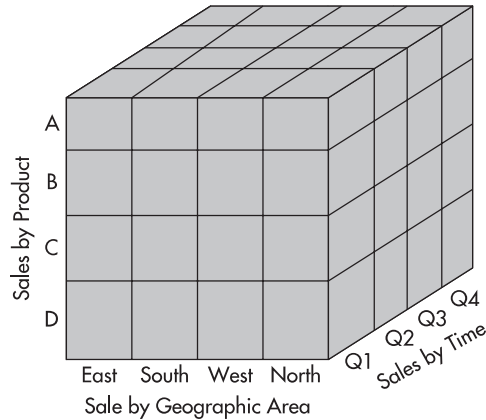


Figure 1-1 OLAP multidimensional cube

geographic area, and time (date). Cubes are also known as star schemas, which is a type of data model. The term “data models” refers to how the data is designed, which may be star, snowflake, third normal form, or other style. These data models all present data in various formats: physical-type models consist of tables and columns, while logical-type models consist of entities and attributes. These star schemas are also known as data marts or analysis areas. Note that while many consider data marts to be cubes, in reality they can be other types of designs as well, such as third normal form. More information on data model types will be presented as we progress through the book.

Any combination of simple dimensions can produce thousands of reports at the click of a button. Drill down from country to province, then from calendar year to calendar quarter to calendar month. Further drill-down to major city and days is also possible if the underlying structures allow for the data at such granularities. Drill-up is the same concept as drill-down but going from lowest to high level of granularity (month to quarter, then quarter to year). Figures 1-2 through 1-5 provide an example of drill-down using two of the three dimensions: geographic area and time period.

An important point to mention is that the drill-down or drill-up is typically quite quick. Response time from clicking the button in the BI tool to the presentation of the report should be seconds to minutes. Remember, this is not an operational

Canadian Sales		
		2007
Geographic Area	Totals	
Canada	945	945

Figure 1-2 High-level geographic area and year report

Canadian Sales		
		2007
Geographic Area	Totals	
BC	36	36
Alberta	131	131
Saskatchewan	173	173
Manitoba	36	36
Ontario	67	67
Quebec	251	251
New Brunswick	36	36
Nova Scotia	78	78
PEI	86	86
Newfoundland	51	51
Canada	945	945

Figure 1-3 Drill-down by geographic area showing provinces

transaction system where response times are subseconds. In the OLAP world, since there can be millions and possibly billions of underlying data rows, response times should be quick but can span a number of minutes. An important aspect of business intelligence is performance. Results must be usable and therefore are expected to not take one to two hours each time a button is clicked.

Dashboards and scorecards, as shown in Figure 1-6, are another special type of reporting focused on visual representation. These typically contain highly aggregated key performance indicators showing how business measures have been doing and how they are currently doing against some predetermined range. A scorecard is the

Canadian Sales					
		2007			
Geographic Area	Totals	Q1	Q2	Q3	Q4
BC	36	12	6	12	6
Alberta	131	15	13	20	83
Saskatchewan	173	90	22	13	48
Manitoba	36	12	6	12	6
Ontario	67	10	16	38	3
Quebec	251	50	102	87	12
New Brunswick	36	12	6	12	6
Nova Scotia	78	38	10	25	5
PEI	86	6	10	46	24
Newfoundland	51	12	24	9	6
Canada	945	257	215	274	199

Figure 1-4 Drill-down further on date from year to quarters by province

Canadian Sales													
		2007											
Geographic Area	Totals	Q1			Q2			Q3			Q4		
		Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
BC	36	3	4	5	1	2	3	4	5	3	2	1	3
Alberta	131	4	5	6	4	3	6	8	9	3	32	45	6
Saskatchewan	173	2	43	45	6	7	9		7	6	4	32	12
Manitoba	36	3	4	5	1	2	3	4	5	3	2	1	3
Ontario	67	2	4	4	5	5	6	4	32	2	1	1	1
Quebec	251	2	43	5	6	87	9	0	0	87	6	1	5
New Brunswick	36	3	4	5	1	2	3	4	5	3	2	1	3
Nova Scotia	78	32	4	2		4	6	3	21	1	3	1	1
PEI	86	2	3	1	4	3	3	5	34	7	8	9	7
Newfoundland	51	3	4	5	1	2	21	1	5	3	2	1	3
Canada	945	257			215			274			199		
		56	118	83	29	117	69	33	123	118	62	93	44

Figure 1-5 Drill-down further on date from quarter to months, still by province

same concept as a school report card. The scorecard shows how specific business key performance indicators have been doing in relation to previous levels. The dashboard is similar to an automobile dashboard with visualization of how the enterprise is currently operating. These reports are typically for upper management to get a feel of the organization at a high level. But there’s no reason why these types of visual reporting mechanisms cannot be used in the daily operations of the organization, for instance in a call center scenario; perhaps showing number of calls received or average time spent per call compared to average per week and average per month.

BI Tool and Architecture

Remember the first lines from the BI definition section of this chapter: “Business intelligence is an umbrella term referring to skills, processes, technologies, applications, and practices used to support business decision making.” Well, when you’re looking at BI tools, do not simply look at the user presentation perspective. BI spans the data warehouse system as well as focusing on the retrieval and presentation aspect of data in a specific context.

A major component of BI is how the data is captured, cleansed, and held, and the compatibility of the data with the user presentation tool. The underlying data repository and the BI tool work hand in hand. Both must be designed to function as one, or else there will be a number of disconnects. The idea is that the data must be architected for retention as well as for usage, which means that the BI tool must be able to handle special designs in an efficient and timely manner. If the tool requires the data in one particular design other than what has been supplied, more effort will surely be required along with additional data movement and optimization.

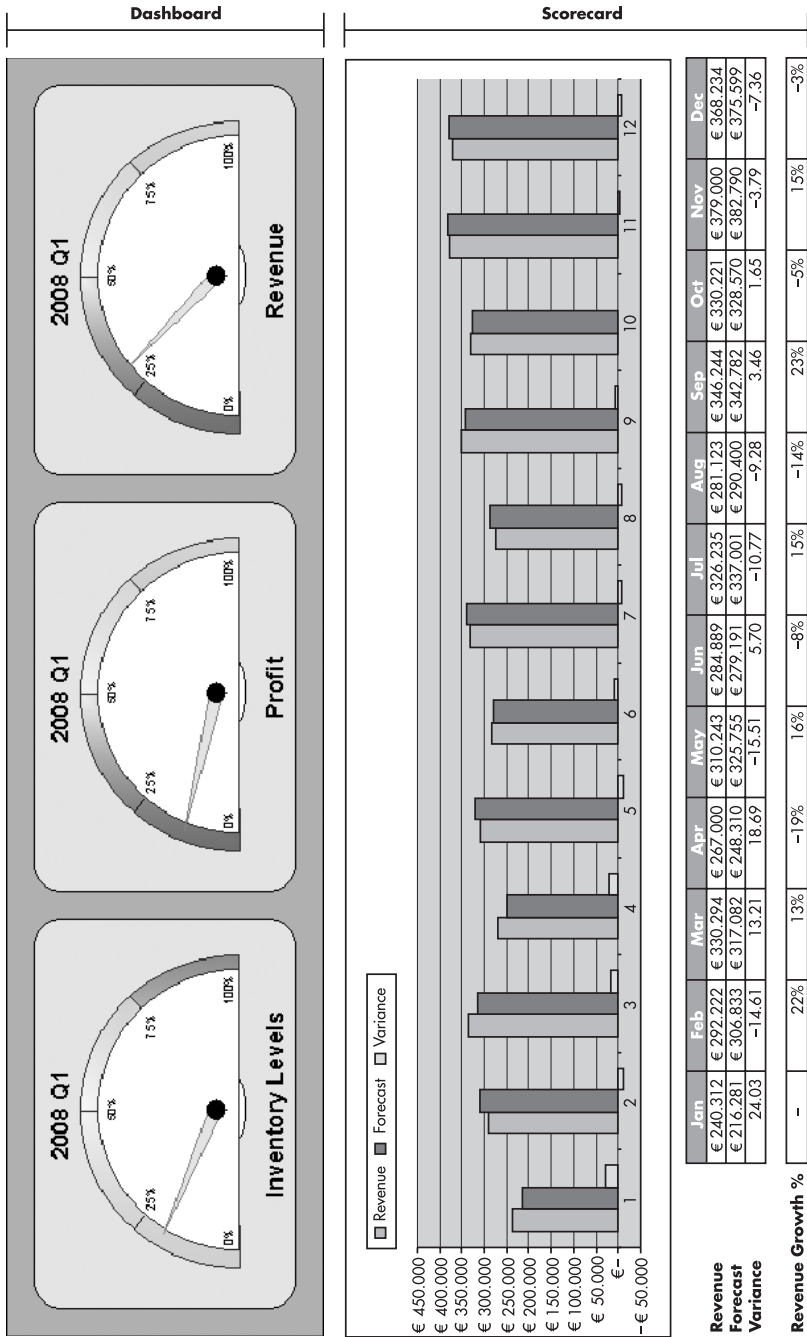


Figure 1-6 Dashboard and scorecard

In short, the BI tool is not an independent item. It must be coordinated with the underlying database, architecture, and overall solution. Business intelligence is a solution, not just a tool or cube or specific report. There are many vendors selling tools, each with their own nuances and special usage scenarios. Be sure the tool is compatible with your databases, your platform operating system, browser, and planned technical architecture. Test the tool with your own data to ensure that it functions as advertised. Sales pitches can be quite slick; test all features on your own system before spending thousand of dollars.

Advancements Due to Globalization

Over the past several decades, with technology advancements and modernization leading to information being much more easily accessible on the Web and with experts and consultants travelling the world more frequently, there have been global advancements in BI and data warehousing. The diverse world cultures and different business structures and strategies have ironically helped the development of standardizations. Telecommunications companies in Jakarta, Indonesia; Mumbai, India; or Toronto, Canada can all reap the benefits of globalization due to the use of best practices, line-of-business or industry data models, tools, and development and/or usage methodologies.

Bill Inmon is accredited with developing a standardized architecture methodology for structuring enterprise data to support business intelligence. Ralph Kimball also developed an architecture methodology consisting of guidelines and components for specifically building business intelligence environments based on a data bus structure. Both are keen to focus on business requirements, whether data and/or reporting, to help the business attain a higher level of business decision making. Individually or together in a hybrid fashion, both methodologies form approaches in designing and building an environment to help organizations optimize their performance in managing their business decision-making process, which is called a “data warehouse.”

Business intelligence is a component of the data warehouse. Understanding a business intelligence environment requires an appreciation for the overall data warehouse system and for the different approaches to designing and building such an environment.

Data Warehouse Overview

Business intelligence was discussed prior to the data warehouse topic because most data warehouse systems come into existence to create some sort of business value, which is typically business intelligence. Now let us look at the definition of a data warehouse in more detail.

Definition

A data warehouse (DW), aka warehouse, is a system for collecting, organizing, holding, and sharing historical data. It consists of “used” data as the data comes from operational systems that capture and use the data within the context of that system’s purpose. Of course, other systems or sources are also possible, but in a DW project the term “operational systems” is widely used. There is usually more than one source system for a data warehouse. Data warehouses are typically thought of as enterprise-wide, but in many instances can be focused on a particular line of business such as finance or marketing.

The term “data warehouse” is often used to refer to a data warehouse system and at times in reference to the data warehouse repository. Throughout this book it will be used to refer to the overall system. The term “data warehouse repository” will be used when referring to the large central database or its design, which are components of the data warehouse system.

A data warehouse is used by the business users for decision support. Decision support in this context is synonymous with business intelligence, which is the usage of the data and the manner in which it is gathered, held, and presented within the data warehouse.

Business users, aka end users, run queries in one manner or another on the data within the data warehouse environment to support their process of making business decisions. The type of inquiries can range from simple queries, trend analysis for data over time, comparative analysis, data mining for associative analysis, extrapolation or predictive analysis for future expectation, and mixtures of these or others depending on the business usage requirements.

Many confuse a data warehouse system with a data warehouse project. Look at it this way: A project, regardless of the line of business or its purpose, has definitive start and end dates. A data warehouse, on the other hand, is a system that has lifecycles, as shown in Figure 1-7. A data warehouse is built or added to via projects.

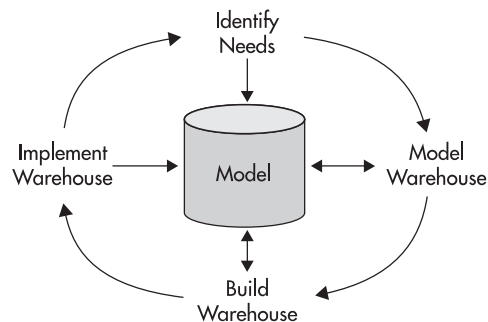


Figure 1-7 *Data warehouse lifecycle*

For instance, to create a data warehouse, a project is undertaken to determine what exactly to build. Another project at a later time may be funded to add or expand the data warehouse environment yet again.

A classic data warehouse lifecycle is to identify the business needs, or requirements, as well as high-level technical requirements. Remember, a data warehouse has both a business and IT symbiotic relationship; one cannot survive without the other. Once it is determined why the data warehouse effort is required and a budget has been approved with the go-ahead to build, the design or modeling phase (or macro phase) begins, which includes technical and data architecture, data modeling, process modeling, and so forth. Then after the planning and designs are done, the actual building can begin. This step involves the micro-tasks of physically building the environment, which includes creating the database, determining indexes, writing extract, transform, and load (ETL) jobs, writing reports, and so forth. Lastly, for the project at hand, the implementation of the warehouse is the setting of all the components into a production status and the deployment to the business user community in order to actively expand the business insights and opportunities. As a disclaimer, these are the high-level steps in a data warehouse project effort, but that is not to say that the underlying tasks must necessarily be in a waterfall development approach.

Data Warehouse System

As with any system, the main components are input, process, output and feedback, as shown in Figure 1-8. For a data warehouse system, the input involves identifying and capturing the data. Data quality at this point is critical because any incorrect data will cause inaccurate output as it trickles down into all underlying processes,

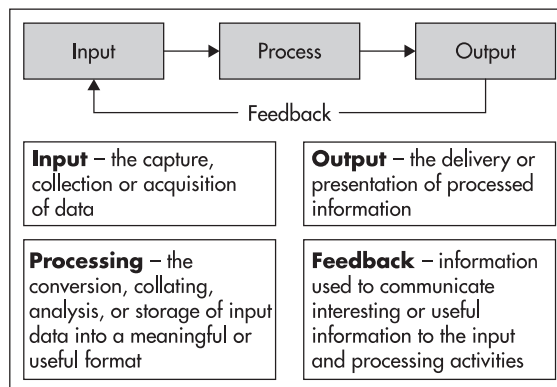


Figure 1-8 Basic system components

subsystems, and eventual business analysis and decisions. Transforming and loading the data into a central environment can span the input and process areas.

The central environment is typically one large database but can be a combination of databases possibly on different servers in different locations, but regardless of the setup, they should all be choreographed from one central design. The process aspect of this central area is to transform and hold the data in a structured and organized manner. Structuring the data is done via data architecture and more precisely a data model. A logical data model is used to design and understand the fundamental, descriptive, and associative characteristics of the data and therefore the business. A physical data model is used to optimize the data for usage in a database and related environment.

The outbound portion of the system is to transfer the data to those who need it. The data can be designed in a number of different ways. Typically this is referred to as a combination of data marts, which can span the performance layer and the user presentation area/layer depending on how they are discussed. Data marts are smaller versions of the central environment that conform to the logical and/or physical central data models but are optimized for end user–specific usage. In this context the data marts are referred to as the *performance layer*. This might involve materialized query tables created by the system, specialized access paths, virtual views, and so forth. The user presentation area is the portion of a data warehouse the business people actually use, which is typically done with some sort of reporting tool such as Cognos, Microstrategy, Business Objects, Crystal Reports, and/or SAS. Other tools include Microsoft Access, manual SQL, Excel, or a combination of any of these and others. The user presentation area sits atop the underlying specific data marts.

The feedback portion of the system is based on the output and input portions. When deriving or aggregating data, the result could be required for later use. In this case, it may be practical to keep the aggregation for later usage such as calculations. Therefore this output data is now required to circle back as input into the data warehouse system. This newly derived measure is now sourced from the data warehouse itself and can be used as the foundation of future inquiries. There are certainly methods of optimizing this feedback approach, which all depend on where within the output process it is created.

Data Warehouse Architecture

The architecture of a data warehouse is the design of the data warehouse system. Think of the architecture as the blueprints. One very popular manner of representing the data warehouse architecture is by using a data flow diagram, as shown in Figure 1-9. The reason this is so popular is that it gives a really good overview of the underlying components. A data warehouse is a complicated system. It may seem like an easy

undertaking—take source data, centralize into a nicely structured database, and run some reports—but in reality it is quite complicated and time-consuming.

I have seen naïve projects begin without regard for business usage and attempt to build a data warehouse without business purpose, without source system insights, without proper planning, and without executive support. I’ve also seen projects in which an IT data manager decides to build a data warehouse in hopes that business will jump on its existence once it is built. One particular IT-focused customer had executive support because they talked upper management into the amazing possibilities of having a great understanding of the organization’s data. Well, upper management missed one important point—tangible usage. The project seemed to be lasting forever and could no longer be contained solely in the IT data management budget because resources were being consumed that affected other budgeted efforts. A data warehouse architecture data flow diagram was eventually created and it became apparent that efforts were only focused on the data population and organization aspects of the overall solution. If you build a portion of a product simply because it is a good idea, it does not necessarily ensure that someone will purchase it. There are costs to continue development, costs in advertising/promoting, costs in distribution and usage. Ensure that the business requires and will use the product before building it, no matter how great the idea or concept may seem.

Figure 1-9 shows the data flowing from left to right within a data warehouse system. Used alone, this graphic helps in understanding the components of a data warehouse system. Adding organization-specific information to it will help communicate a realistic view of the overall solution and give insight into ongoing efforts.

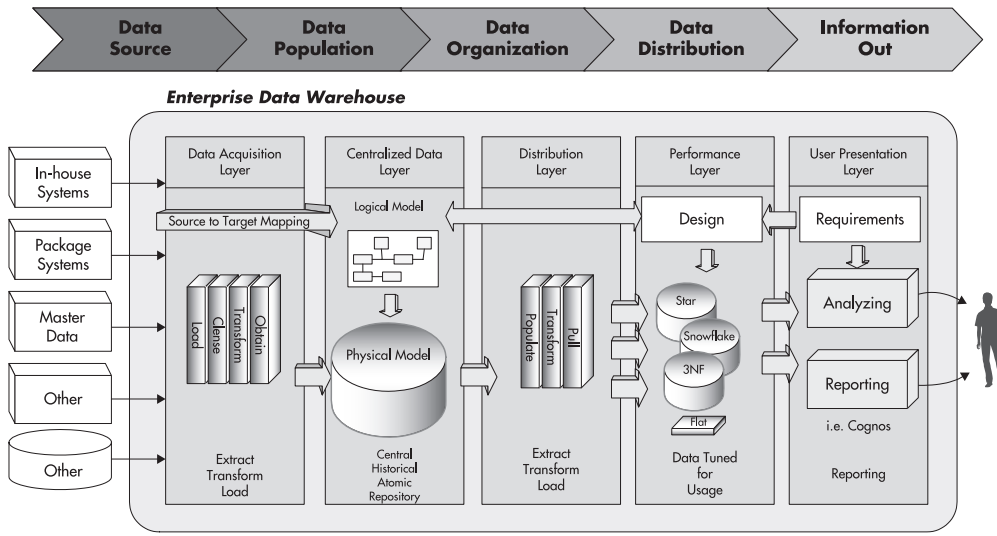


Figure 1-9 Data warehouse architecture—data flow

At the left are the input source systems, which are typically operational systems within the organization and can include feedback sources from nearly every one of the other layers. Following are numerous processes to capture, populate, and organize the data. And finally on the right is the distribution of data and the usage or output of the information to the business community.

Each one of these layers should be expanded in the design, build, and implementation phases of the project, but first, requirements must be gathered. The *raison d'être* (French for “reason for being”) for a usage must be determined. If a budget is to be spent for the development of a data warehouse and therefore a monetary value placed on the system or product, there should be a real, tangible useful purpose to its existence with an estimated return on investment (ROI), as with any business venture effort.

Some argue that structuring the data is of massive value to the organization in itself, which is true, but without a business-specific usage, how do you know if spending two weeks on one particular data aspect is of later value or not over some other data concept? In general, do not build a pure data foundation in hopes of business becoming interested at some later point; build with business value goal or purpose up front. An organized and structured data foundation should be a by-product of the business value approach and solution.

Data Flow Terminology

Many in the data warehouse world use the terms “top-down” and “bottom-up” when referring to a data warehouse. Most discussions using these terms take them from the data flow architecture. Here’s the trick; if you turn Figure 1-9 clockwise 90 degrees, the inbound data portion becomes the “top” of the data warehouse architecture and the output portion, aka the business usage portion, ends up at the “bottom” of the architecture, as shown in Figure 1-10. Another way to think of data flow is whether the data is entering or exiting the system.

Bill Inmon’s methodology primarily deals with a data warehouse from the top down, meaning from the data point of view but not to say without business purpose. Ralph Kimball’s methodology is from the bottom up, meaning business purpose above all else with data to support it. The author finds a hybrid approach very effective, and it promotes efficiency in project and efforts. Both methods are really driving to the same end point, which is to support the business being able to make informed decisions by structuring and organizing the underlying data.

Many confuse these top-down and bottom-up terminologies, but as long as they are distinguished in the discussion, all is fine. The reason for the confusion is that many misunderstand the context of the discussion. For data flow the top is the data and the bottom is the usage. For information usage, the top is the business usage, typically business reporting, and the bottom is the supporting data foundation, as seen in Figure 1-11. Both views are appropriate as long as the context is

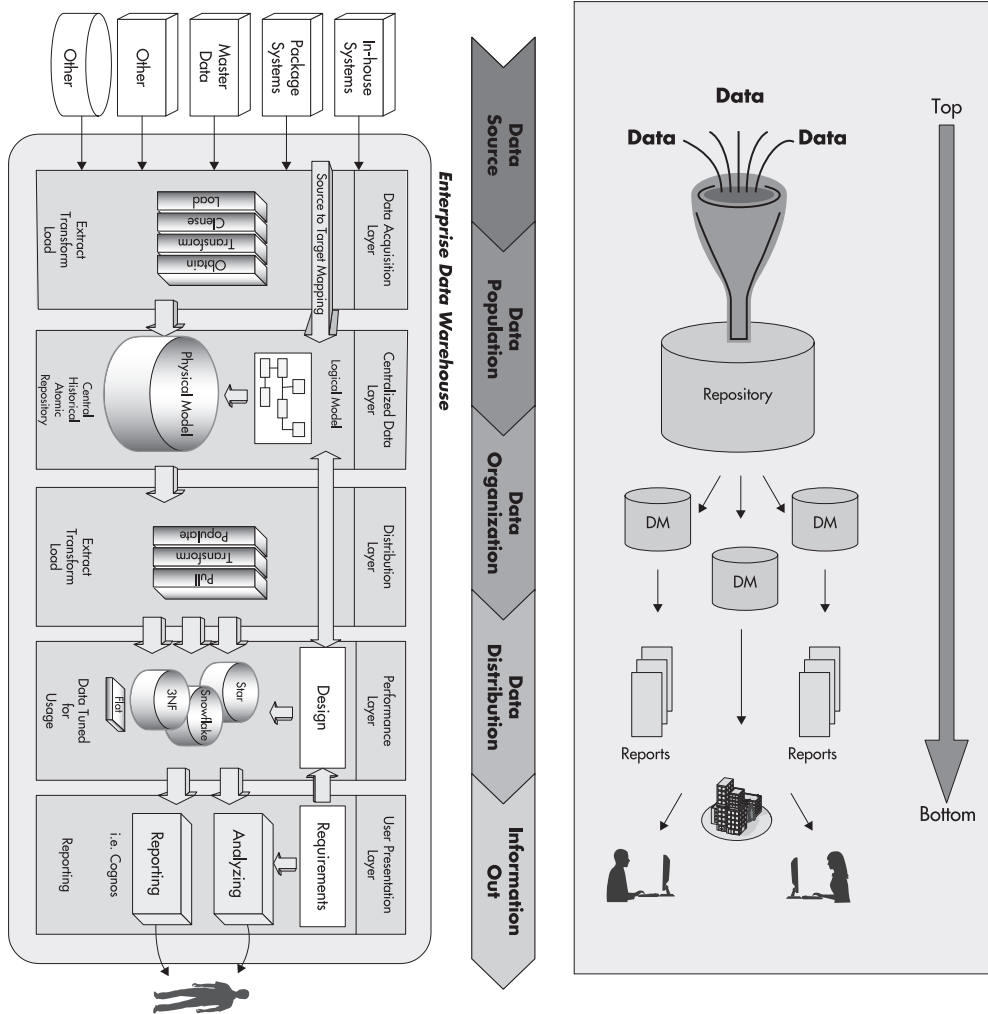


Figure 1-10 Data warehouse architecture: data flow

distinguished and maintained throughout the discussion. This book will always refer to the data flow top-down and bottom-up scenario unless otherwise noted.

Data Warehouse Purpose

A data warehouse environment is built to hold historical data integrated from a number of source systems in an organized manner. Operational systems are built for specific functions, such as point-of-sale processing, billing systems, inventory control, and so forth. These systems are not always enterprise-based and are not built

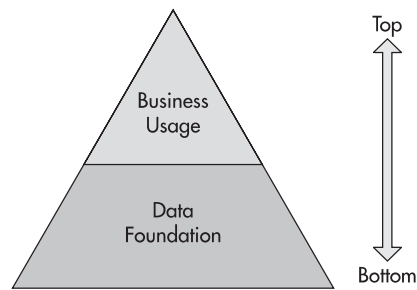


Figure 1-11 *Data warehouse architecture: information usage*

for data analytics or data mining. Hence a new environment must be created to merge the data from these systems into one central area, called a data warehouse system, for overall enterprise usage.

Data quality can be an issue in one system, but when merging disparate systems, data quality is paramount. I have never seen an organization with absolutely perfect data. One customer came close but then they only had one source system. Once they merged a second system into their data warehouse, they were surprised to find that data quality issues rapidly surfaced.

Due to the merging of multiple systems, a data warehouse must pay attention to modeling, aka structuring or organizing, the data to ensure a common vocabulary and flexible design. For instance, system A has a sex code of 0 for Male and 1 for Female. System B has a gender code of M, F, U (unknown). When merging these two systems, a commonality in vocabulary and data values must be created. We might decide to call this Gender Type and define the codes as M, F, and U representing Male, Female, and Unknown. This seems quite simple for this example, but ask several business users from different departments in your organization what the term demographic means to them. I am sure the result will be a number of different answers.

The primary reason for having a data warehouse is to sort out the “spaghetti” mess that every organization has either from disparate systems built over the years, or merging of organizations, purchased systems, or whatever. The mess is in the terminology, or vocabulary, and in data values. Figure 1-12 shows how operational (source) systems can each play a role in enterprise analytics. Without a centralized environment, reporting becomes difficult, inconsistent, incorrect, of high maintenance, and unreliable due to data quality issues, redundant loading routines, impact to source systems, and so on.

A centralized “atomic” data warehouse repository, Figure 1-13, cleans up most of these issues. If you centralize the data for the enterprise, a vocabulary emerges, allowing the users to all speak with the same terminology. And it allows all data to have a high level of quality since each object must be analyzed and profiled for it to be merged into a central data warehouse repository—or at least should be.

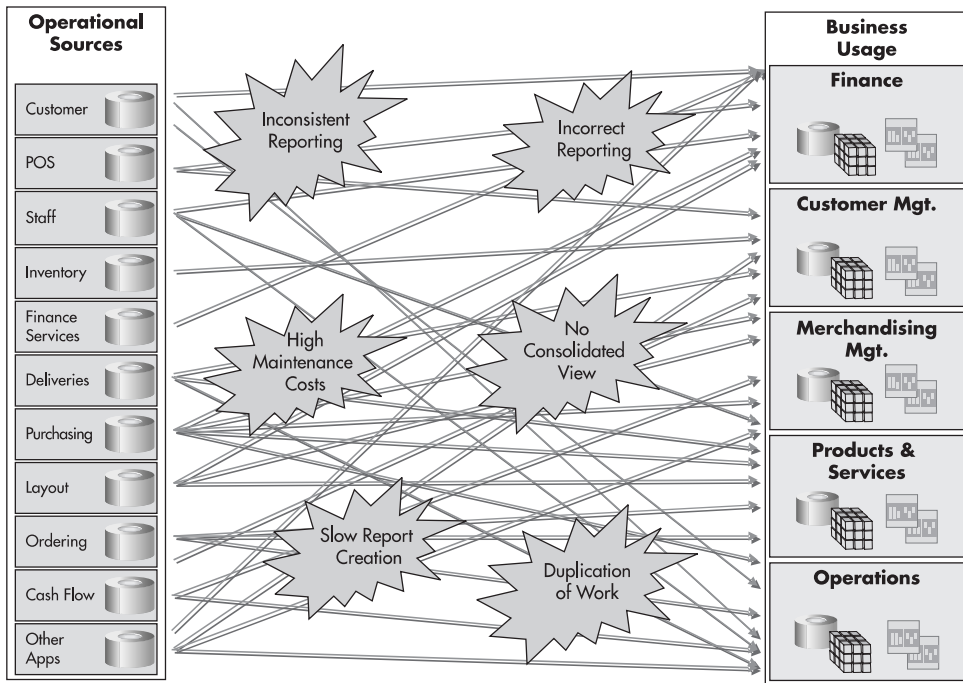


Figure 1-12 Reporting spaghetti mess

With reporting from a central environment, data on each report can be accounted for. If there are any issues, the sourcing of each data component can be traced back to its origin simply because the loading of the data warehouse is under the control of the data warehouse group and hence all loading programs can be reviewed.

The term “atomic” in this sense refers to the level and granularity of the data. The business could use a term such as demographics, but IT would decompose this into data items such as person, age, income, address and so forth. This fine level of data is the atomic data level and the granularity is the level of data capture. The spaghetti mess also refers to the disparate granularities of data being captured.

If a transaction happens several times a day, then the granularity is at the level of data capture, which is the transaction level for this example. For instance, a customer purchases an apple from a fruit store in the morning. In the afternoon the same customer purchases another apple from the same store. The lowest level of granularity is each “purchase,” or point-of-sale transaction.

The atomic data level is the lowest level of possible data capture and at the finest level of data components possible. The data model design is a normalized form,

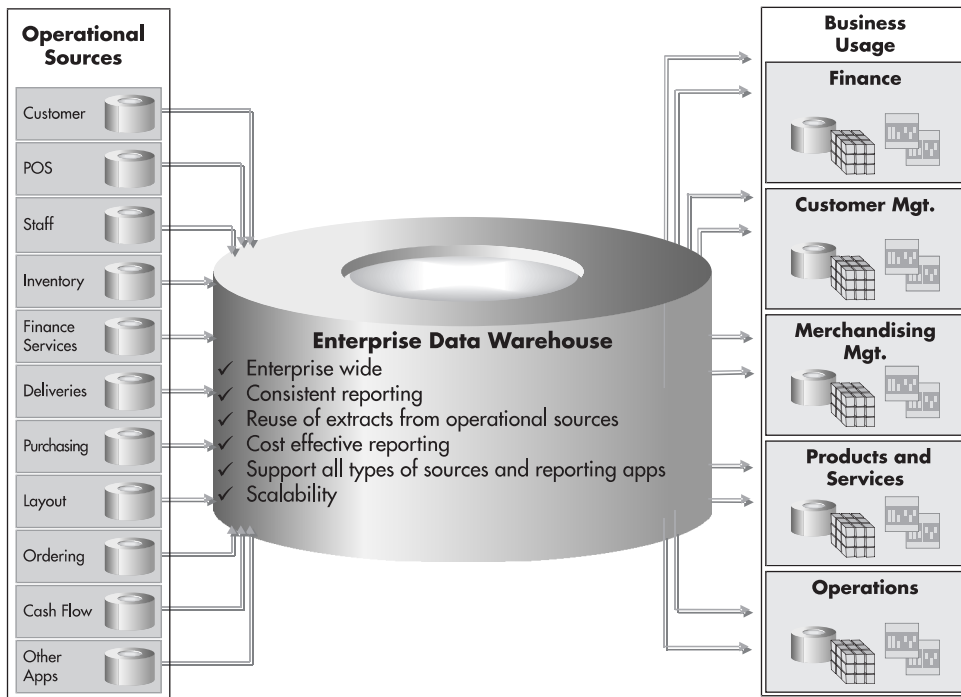


Figure 1-13 Structured and organized data warehouse repository

which could be third normal form, bus architecture, or star schema. At this level the idea is to capture the lowest grain of the business events and data pillars (terms that will be described in Chapter 2).

Data Structure Strategy

Data warehouses typically evolve from a need to understand information at the enterprise level and also from a strategic or tactical direction within the organization.

As organizations grow, especially after mergers and acquisitions, management needs to know what information is available and how data from disparate systems mix together, as shown in Figure 1-14. Or the organization may be keenly aware of the competitive environment and needs an edge to ensure its position within the marketplace. Both scenarios require an in-depth knowledge of the organization's data asset. For this a master data management effort is usually undertaken. This entails identifying all data values and setting a common vocabulary and structures along with ownership, change control, and so forth.

Vocabulary means to set a definition and example, if possible, for each data object. *Structures* refers to creating a blueprint for each object from fundamental,

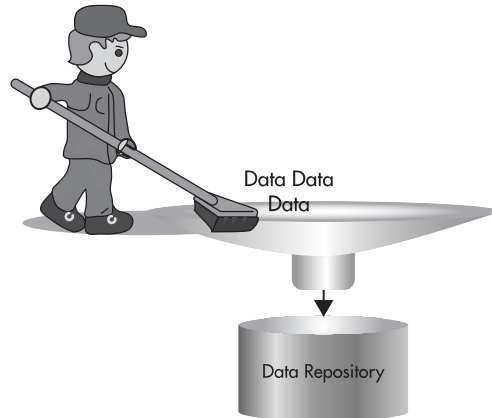


Figure 1-14 Structuring and organization enterprise data

descriptive, and associative viewpoints. Identifying the actual data is exactly what it says; each value for every object should be identified and documented. If gender type can only refer to Male, Female, and Unknown, then this is the universe of values for this data object. Anything new would be something to look into to ensure that there were no processing or loading errors. You may not want to profile an individual's last name, but you would certainly want to remove pre- and post-titles such as Dr. or PhD, as these are data objects themselves but are only necessary if they have value to the business.

When you're selling your home, buyers are much more impressed if the place is neat and clean; the same idea applies to an organization. Having all your data structured and organized lets a potential investor and/or buyer know that management is on their toes, and it is a requirement for regulatory compliance in certain countries. It also gives organizations insight into the operations, thus allowing management the opportunity to make well-founded business decisions at all planning levels. A structured and organized data environment can give management an insight into the worth and value of the company and the company's placement within the marketplace.

Data Warehouse Business

Personally, I like to think of a data warehouse system including its management and development teams as its own independent company within the larger picture of the organization. In this context the company manages a data warehouse environment and produces customer solutions, which include business intelligence solutions. Customers commission the company to build tangible optimized data environments, typically with business reports specific to the customers' needs.

Customers are the business users, and commissioning is a service-level agreement to ensure timely and accurate data within the data warehouse system along with the creation of data marts and business reports for the customers. Data quality is expected or the customer goes elsewhere, just as in an open market system.

The reason for thinking this way about the data warehouse system is that it is quite a big system within the organization. The fundamental aspect is data, which is a major asset of the organization. Thinking of the whole thing as a business allows the involved people to consistently strive to bring in more revenue, just as in a real business. Individuals no longer take on the attitude of an employee of a large organization, but that of innovative entrepreneurs within an up-and-coming business-solution organization. Individuals do not run off and build quick solutions for specific customers; they discuss requirements and issues with management, who determine the appropriate course of action for the data warehouse organization. Each new project, whether to enhance the current system or expand the system for new customer requirements, is an extension of the organization and of product output.

For these reasons and this perspective, a data warehouse is a true organization-strategic asset and therefore should have executive ownership. This means a management focus on enhancing the asset, structuring the asset to meet the organization's strategic direction and ensuring data quality, not only within the data warehouse system but for the source operational systems for data pertaining to the data warehouse. Centers of Competence or Excellence are good examples of how to run the data warehouse and business intelligence division within an organization.

Frequently Asked Questions

I've been involved in more data warehouse projects than I can count all over the world, and it seems as if the same questions are asked over and over again by the development team, the business users, and executive management alike. It's interesting how globalization has people in the four corners of the earth from different cultures, different languages, and different lines of business all asking the same types of questions, as described in the next few sections.

A data warehouse system can be quite the undertaking, with many intricate processes, designs, requirements, and the coordination among all of these. All these technicalities, all the while trying to manage the organization's internal politics, can be daunting. After spending time deeply involved in the details, it can be quite easy to lose sight of the fundamentals and exactly why the strategy and development have taken the direction they have. Even after organizations have built previous data warehouses, they can still become disenchanted with an enterprise data warehouse effort or business intelligence altogether. For these people, the following few sections are a compilation of classic recurring questions and (hopefully) satisfying replies.

Feel free to send your questions to the Data Warehouse Mentor group on www.linkedin.com. I or another member will gladly respond to the best of our abilities to help you out in the most timely and accurate manner in which we can hopefully give valuable and usable information in your context.

Current Systems Good Enough?

Rather than creating a new data warehouse system, why not just access the source (operational) systems directly?

- ▶ Typically, operational systems are not interested in keeping history on all the data. Therefore a new data warehouse environment must be created to ensure that all versions of the data are captured over time.

My source system does capture and hold all data over time, so why create a data warehouse?

- ▶ Data in source systems is optimized for usage. If the source is from a point-of-sale system, a mediation system, or a billing system, for example, these systems are specifically designed for their primary purpose and not for associating one component to another in an analytically desired manner. To do this the data must be pulled, structured, and placed into a data warehouse system so it can be used in whatever manner desired. This is also why capturing the data at its finest level of granularity is rather important.
- ▶ Also many source systems are closed. Pre-purchased packages usually have no documentation about the underlying tables, columns, or how they associate to each other. This makes understanding the context of the data nearly impossible and querying the data very difficult. The software manufacturer normally makes the data available through the software application or via an application program interface (API).

My source system is open, has full documentation, and holds all data over time. Why should I build a data warehouse?

- ▶ Well, if this is the case, then perhaps you do not require a data warehouse for your current purpose. Check if you have data quality issues; fix if necessary and query your data as is—if you don't mind having large and long-running queries on your operational systems.
- ▶ Remember, your operational systems run the company business. If you interfere with these by slowing down operational response times or delaying business processes or backups, this can hinder your business productivity. It's best to create a new environment and copy the data from the source system as required.

- ▶ Typically more than one source system is used in business intelligence, and a data warehouse environment is perfect for merging such disparate systems and aggregating the overall data, allowing queries to run on one common enterprise environment.
- ▶ Another important point is in setting an enterprise vocabulary for your organization. With multiple source systems and applications, it is very likely that a term in one system is similar to but still different in another system. Or it may be completely different between both systems. Having a central data warehouse environment allows for a master data management effort in terminology and data values as well.

Well, all this sounds great; why not replace our current operational systems with a data warehouse?

The answer goes back to several of the same points mentioned earlier:

- ▶ Data in operational systems is optimized for usage. Point-of-sale systems, mediation systems, or billing systems, for example, are all specifically designed to capture specific real-time information.
- ▶ The model and therefore database for a data warehouse is designed specifically for historical data in a manner optimized for holding and reporting. Operational systems are designed for capturing and subsecond response times. Data warehouses can have millions and billions of rows gathered over the years; therefore, response times can be minutes or even longer depending on the requirements.
- ▶ Remember, your operational systems run the company business. If you interfere with these by slowing down operational response times or delaying business processes or backups, this can hinder your business productivity. It's best to create a new environment and copy the data from the source system as required.

What Is the Value of a Data Warehouse?

Why create a data warehouse effort? Why not just obtain (referring to their own departments) a handful of smart IT folks to create the reports as need be?

Many organizations have business departments with their own IT staff who are maintaining certain aspects of their operational systems and/or are tasked with producing their required reports and/or being the go-to person to produce on-the-spot reports based on specific requirements. These IT folks are constantly trying to get access to one system or another, copying data from one source system to their own environment and replicating that data multiple times as it is sent all over the department. The end result is that different groups have different concepts, vocabulary becomes local rather than enterprise, and reporting becomes misleading.

The point is that the value lies in having a central common enterprise area for all business users to access the same underlying data. The context of that data or information can be in any manner the individual users determine appropriate for their duties and purposes. But the underlying data would be common to everyone. This means the vocabulary is common, the data values are common, and the structures are common to all in the organization. Data becomes an asset to the organization as a whole, not just departmentally. Banks, insurance firms, phone companies, and airlines cannot operate today without computer systems that can handle the underlying mountains of data. If the data is incorrect, the results are incorrect, which leads to inappropriate decision making (for those decisions based on the underlying data, of course). The point is to ensure that the fundamental asset is secure and trustworthy.

Several years ago while I was in discussions at a customer site in Norway, both marketing and finance department representatives were sitting at the same table. I thought, what a great opportunity to demonstrate the business value of a data warehouse. I asked how many products the organization produces. The answer from one department was 450; the other department quickly jumped and literally said, “Are you nuts? We have 14,000 products.” The point was made; they each had their own data and interpreted it in their own way. The value of a data warehouse is to ensure that the organization can base their decisions on the same fundamental data, using the same vocabulary/definition throughout the enterprise.

It is all about ensuring that the enterprise data asset is managed appropriately for the business to access and used as need be.

Actual value in monetary terms is different for each organization, but the business goals are typically the same:

- ▶ Deeper insight into product base including quality assessments
- ▶ Deeper understanding of current business processes
- ▶ Deeper insights into customers and customer relationships
- ▶ In-depth knowledge of current operations
- ▶ Identification of market opportunities
- ▶ Improved marketing strategies
- ▶ Deeper insights into financial areas such as
 - ▶ Customer accounts and trends
 - ▶ General ledgers
 - ▶ Product costs and profits
 - ▶ Transactional analysis
- ▶ Comparative insights against the competitive environment

How Much Will It Cost?

This is always a difficult question. Cost is relative and depends on the current environment and what the organization wants to do with a data warehouse. There is no straight answer to this question. To access the cost, consider the following areas:

- ▶ Current technology vs. anticipated technical requirements
 - ▶ Servers or machines
 - ▶ Databases
 - ▶ Disk space
 - ▶ Data models
 - ▶ Tools:
 - ▶ Data capture tools
 - ▶ ETL tools
 - ▶ BI usage tools
 - ▶ User licenses
 - ▶ Maintenance costs
- ▶ Current vs. expected expertise
 - ▶ Is appropriate talent available in-house?
 - ▶ Are resources currently available?
 - ▶ What are the education costs for IT and business users?
 - ▶ What are the external expert resource costs?
- ▶ The scope for the project
 - ▶ Top-down or bottom-up
 - ▶ Finer focus is more effective and cheaper
- ▶ Ongoing system costs
 - ▶ Production maintenance

When all is said and done, cost will be a factor of expected return on investment of the overall data warehouse and business intelligence system.

How Long Will It Take?

This is another good question that keeps most data warehouse managers awake at night. Many projects are given a specific timeline with a drop-dead date to finish all

development and produce a final usable result. The problem is that most data warehouse projects succumb to scope creep. More and more seemingly small items are constantly added to the project as it progresses. These may seem small and inconsequential at the time, but cumulatively they do add up to increase the project deliverables and ultimately will extend the deadline. Later, when difficulties arise, management has a funny way of jumping back to the start and stating how everything was supposed to be done in xx months. The little extensions seemed trivial, but somehow all hell broke loose and the project turned into a runaway state.

Best practice is to limit initial scope. Focus on fundamental data before complex derivations (cost and profit for instance). Do not model the entire organization but only the most fundamental data as required by the very clear business purpose or goal. Ensure that the project effort has a very precise documented and approved plan with set goals. A qualitative goal is nice, but it must be quantified so it can be measured and attained.

Best practice is to not start from scratch. A good practice is to purchase as need be. For example, purchasing a prebuilt data model will help greatly with organizing and structuring the enterprise data with naming standards, definitions, and relationships between the components. However, be warned that purchasing an intricate data model must be done with an understanding that the model may need tweaking and should be used as a reference. No purchased data model is exactly what any one organization is looking for, but with expertise it can certainly ramp up efforts on any data warehouse project. Depending on an initial scoping, and given appropriate client subject-matter experts and source-system analyst insights, a purchased model should take between six to nine weeks to map out the organization's fundamental data.

Best practice is also to ensure that the project has a seasoned full-time data warehouse project manager. Just because a person is a good project manager for an operational system project does not imply that this person is a good data warehouse project manager. And ensure that the project is not run by the timekeeper, a person good at driving a project plan but little or no authority or experience in managing the actual project and resources. Ensuring that a task is done does not make a project manager. Knowledge, experience, and authority are key to the project manager role.

Best practice is to ensure that a data warehouse project has a seasoned data warehouse architect. A project manager removes political hurdles and plans the steps of the project, while the data warehouse architect ensures that technical details are correct and data flow is proper and development is headed in the proper direction. For that matter, ensure that the data warehouse has senior experienced data warehouse and business intelligence staff. Do not populate the staff with rookies, as this will greatly slow the project down.

With a fully seasoned staff, a limited data warehouse effort with a kick start (not from scratch) should take six months for the initial stage. So the next question is, what is a limited scope?

With one particular customer, I was tasked to determine how cost was calculated for a specific line of products and how that calculation was used in a particular department. The idea was that cost is an essential concept to the business and needed to be fully understood in that department.

Usually when cost or profit comes into a data warehouse scope, it becomes quite a complicated matter. While an organization is very keen to understand these two areas, especially for analysis, these are very difficult and intense concepts within any organization, with a large trickle-down effect to many other departments and areas within the organization. My suggestion is to not begin a data warehouse initiative to determine cost or profit up front. First, start the data warehouse based on fundamental data; get all your ducks in a row before determining how they fit together and form cost or profit insights. As a rule, business concepts such as cost and profit must be defined by the business as input to the data warehouse system. All terms and calculations must be known to the data warehouse up front; this is a major aspect of determining an enterprise vocabulary.

In trying to determine cost for this one customer, the business analysts were brought in for an understanding of how the business functions in that area. It turned out that they pulled certain types of data from several operational systems and created a special report with underlying derived data. This was then passed over to another person in the department, who imported it into their MS Access database for further massaging. Later the secretary imported the data into Excel and distributed the file to at least ten other individuals, who proceeded to derive what they required in their own special way. The problem was that when one of those users discussed cost with the business analysts, they had their own twist on it, and none of them had a holistic view of the concept.

The point is to be aware of the implications of the project focus before determining the duration of the project effort. This can be quite difficult but an up-front necessity before planning on overall duration.

What Will Make Us Successful?

This is a great question, which every management person should be asking before the project begins and every month thereafter. How can we succeed with this project, with these resources, with this budget, and with this timeline?

Scenario: You win the lottery and end up purchasing an existing company. A vice president approaches you and says we need a data warehouse. Okay, you say, that sounds interesting, let's discuss it. The first thing is to ask the guy for a definition of a data warehouse and how will it help the business. The VP is just beaming with enthusiasm and explains that he heard that all the data in the company should be pulled into one central area, and once this is shown to the business, they will just love it and they will do lots of amazing things with it, which will advance the business by leaps and bounds. What would you say?

Nothing can guarantee that the project will be successful, as there are always risks. To remove or limit the risks, the following best practices have been proven time and time again.

Step 1: Research

Look into what a data warehouse is, what business intelligence is, and how it is used. You do not have to become an expert; you can always hire experts, but become familiar with the topic, key points, and basic vocabulary. One of the reasons this book came into existence is that I was talking with a director of business intelligence who abruptly stopped me in our discussion and with a bit of contempt in her voice said, and I quote, “What’s all this talk about data marts, what’s a data mart? I want a data warehouse.” Needless to say, I was taken aback, wondering how that person obtained the position of BI director. I vowed to one day write a book explaining all the aspects of the topic. So the first step is research; go to the bookstore, buy a book on data warehousing and business intelligence (this book, of course), get a pot of coffee, and begin the education process. Key players such as BI director, DW/BI sponsor, and business leads should all be somewhat familiar with a data warehouse system and business intelligence in general.

Step 2: Strategic Alignment

By now the concepts of data warehouse and business intelligence should be solidifying. You are not an expert, but you should be able to discuss the topic. The next step is to determine whether a data warehouse (and this refers to cost, effort, and value) can be useful to the organization. What is the business strategy for the next five years? How is it to be attained? What are the main business process areas and the plan for these? What/where are the current and expected trouble spots? How can a business intelligence effort help out and fit into the overall strategy of the organization?

Step 3: Focus, or Limited Scope

This is an extremely important point. Do not try to boil the ocean. In other words, do not plan on doing everything at once. Focus on something concrete that is important to the long-term strategy of the organization and which will clearly add value to the business. High visibility is critical to have buy-in from the business. But limit the scope to ensure that the effort is technically possible. Think big but start small. A 90-day effort should show some tangible result with visible value. Pick an area that is straightforward, easily understood, has a clear deliverable, and for which you have the technical skills in-house. Do not build until the plan is solid and details are known.

Step 4: Value

While this may seem sensible, many disregard this aspect of a data warehouse effort. For the first iteration of the project, there may not be much value because a data

warehouse requires a foundation to be built, meaning that the startup or learning curve is high in the first round. However, there must be value to the organization. Show how a centralized product vocabulary is created and agreed to, how customers are identified and centralized, or how data quality is in place that was sorely lacking before. Then show that the next phase adds value based on its dependency on the initial effort. The point is to show tangible value and promote it each step of the way to IT, the business, and executive management.

Step 5: Metrics

For business value to be realized, it must be quantifiable. In other words, it must be tangible, accountable, and numeric in some form. We must be able to count something, or compare something. Saying the data is clean is nice but not enough. It must be quantified; for example, 95 percent of all customers now have valid and usable addresses. The quantitative degree of improvement is vital to make a strong impression with the business and upper management.

Step 6: Goals

Success must be seen by all. IT can say their goals of building a foundation of clean data have been reached, but if the business is unhappy with the final result, then goals are not aligned. There must be a coordination of goals and purpose with IT and the business. Do not expect goals to materialize during the project; determine in advance the specific goals and plan on how they can be attained and agreed upon. Decisions about what is realistic, measurable, achievable, within budget and within the timeline can take months to agree on. Do not leave this item open with hopes that all will come to light during development. Keep users in the loop at all times during development, but even before, keep them in the loop during the planning phase.

Step 7: Executive Support

If there is no clear executive sponsor for the project, walk away. If upper management is not willing to put themselves on the line for this enterprise-level project, the probability of a successful conclusion is very low. An enterprise data warehouse is a strategic asset and requires executive oversight and support. The executive must support and champion the project. There must be enthusiasm reinforcing the development team and the business. At one customer's kick-off project meeting, the recently retired CEO showed up to encourage the group. Sounds odd, but this guy had near-celebrity status and sitting with the team discussing how he was back as the executive champion specifically for the data warehouse effort was a super boost to both camps. Throughout the project he would regularly attend quarterly status meetings and really get involved. His support was paramount to the project and really inspired a level of importance. Political roadblocks were removed at the top levels, and the project was able to move forward without hindrance.

Step 8: Business Sponsor

A data warehouse or business intelligence project is specifically for the business to help in the business decision-making process. If the business is not on board with the project or jumps ship midway, the effort is in real jeopardy. A subject matter expert is an absolute must to ensure that IT understands what and where the focus lies. These are the experts who guide the IT people in building what the business requires and ensuring that the deliverable is usable. Think of these people as the owners of the house you are renovating. If they don't like what you are building, they won't use it.

Step 9: Data Management

A real key to creating a data warehouse system with business intelligence is in structuring the data. This one point can make or break a data warehouse or business intelligence effort. Ensuring that the data is organized at an enterprise level, meaning with a vocabulary and structures, is a fundamental aspect of a data warehouse. Purchasing a prebuilt model can greatly help with this effort.

Step 10: Data Quality

Business intelligence is nothing if the underlying data has little or no integrity. What is the point of creating a system if data quality is not a recognized effort? On several projects, overly keen project managers were willing to ignore data quality in hopes that just building the system would be of more value than the data it held. All these efforts ran aground when the system was used by the business. It's kind of similar to opening a gift only to find that you cannot use it because the batteries were not included.

Step 11: Performance Usage

Once all is said and done, when the business presses the proverbial button, if the response time is too long, the solution will not be used. Ensure that the design takes into account the physical aspects such as data volumetrics, database joins, indexing, and so forth. These all form part of the expected performance levels.

Step 12: Flexible Framework

I once bought ten clothes hangers to stow my ten shirts in the closet. I was quite pleased until I bought another shirt! Ensure that the system that is built can accommodate the next phase. The framework must be flexible enough to build upon at a later date. Remember, a data warehouse is a system, meaning that it will probably be added to, and therefore it needs to be flexible to accommodate additions.

